TWO-PHASE CASE-CONTROL STUDY DESIGN FOR BIOMARKER MEASUREMENTS IN GENE-ENVIRONMENT INTERACTIONS STUDIES

Duncan C. Thomas, *University of Southern California, USA*
Sandrah Eckel, *University of Southern California, USA*
Kiros Berhane, *University of Southern California, USA*

Background and Aims: Two-phase case-control designs have emerged as a cost-efficient way of obtaining additional data by subsampling jointly on exposure and disease, combining the main and substudy data in the analysis. We consider the optimal design of studies of gene-environmental (GxE) interactions mediated through a latent biological process for which a surrogate biomarker is available. In the Southern California Children's Health Study (CHS), air pollution and maternal smoking have been shown to interact with genes in the nitrosative stress pathway (e.g., nitric oxide synthase, *NOS2A*), leading to inflammation and downstream sequellae like asthma. We have measured exhaled nitric oxide (eNO), a marker of the underlying inflammatory process, longitudinally on all CHS participants.
Methods: By theoretical calculations, we derived the optimal sampling fractions for biomarker measurements for a substudy stratified jointly by exposure, genotype, and disease status. For simple environment, genotype, and disease variables, the full likelihood can be used; more generally, semi-parametric maximum likelihood or Horwitz-Thompson estimating equations approaches are needed to avoid estimating many nuisance parameters. We illustrate the approach on CHS data by comparing the asymptotic relative cost-efficiency (ARCE) of analyses of the complete eNO data with those obtained by subsampling.
Results: The optimal design for estimating the GxE interaction term typically entails sampling all exposed carrier cases, some of the other cases, and only a few controls; when biomarker costs are 16x enrollment, exposure, and genotyping costs, the optimal choice yields a 4-fold ARCE compared to measuring all subjects. Similar comparisons for main effects, for continuous biomarker measurements, and a range of model parameters will be shown. For the CHS data, subsampling 37.5% of the subjects jointly on disease and exposure only increased the standard errors by 33%.
Conclusions: Two-phase designs can potentially greatly improve the cost-efficiency of GxE studies involving biomarker measurements.